K-means Clustering Method for the Analysis of Log Data

Prof. Rudresh Shirwaikar[1] and Chaitali Bhandari[2]

[1] Shree Rayeshwar Institute of Engg. And Information Technology (Dept. of Information Technology), Goa, India Email: rudreshshirwaikar@gmail.com

[2] Shree Rayeshwar Institute of Engg. And Information Technology (Dept. of Information Technology), Goa, India Email: chaitalibhandari9@gmail.com

***Abstract*—Clustering analysis method is one of the main analytical methods in data mining; the method of clustering algorithm will influence the clustering results directly. This paper discusses the standard k-means clustering algorithm and analyzes the shortcomings of standard k-means algorithm. This paper also focuses on web usage mining to analyze the data for pattern recognition. With the help of k-means algorithm, pattern is identified.**

***Index terms*— Pattern recognition, web mining, k-means clustering, nearest neighbour, pattern recovery.**

I. Introduction

Clustering problems arise in many different applications, such as data mining, web mining, data compression, pattern recognition pattern classification etc. The notion of what constitutes a good cluster depends on the application and there are many methods for finding clusters subject to various criteria's.

Among clustering formulations that are based minimizing a formal objective function, perhaps the most widely used and studied is k-means clustering. *K*-mean clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets (*k* ď" *n*). See "Eq. (1)".

S = {S_1, S_2, …, S_k} so as to minimize the within-cluster sum of squares (WCSS):

$$\sum_{i=1}^{k} \sum_{x_{j \in Si}} \|X_j - \mu_i\|^2 \qquad (1)$$

Where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

II. Standard Algorithm

The most common algorithm uses iterative refinement technique. Given an initial set of *k* means m1,m2…, the algorithm proceeds by alternatig between two steps:

Assignment Step

Assign each observation to the cluster whose mean is closest to it ("Eq. (2)").

$$Si^{\{t\}} = \{x_p : \|x_p - mi^{\{t\}}\| \le \|x_p - m_j^{\{t\}}\| \text{ for all } 1 \le j \le k\} \qquad (2)$$

Update Step

Calculate the new means to be the centroids of the observations in the new clusters ("Eq. (3)").

$$M_i^{\{i+1\}} = \{ (1/|S_i^{\{t\}}|) \sum X_j \} \qquad (3)$$

The algorithm has converged when the assignments no longer change. Commonly used initialization methods are Forgy and Random Partition. The Forgy method randomly chooses *k* observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points. The Forgy method tends to spread the initial means out, while Random Partition places all of them close to the center of the data set. According to Hamerly, the Random Partition method is generally preferable for algorithms such as the *k*-harmonic means and fuzzy *k*-means.

For expectation maximization and standard *k*-means algorithms, the Forgy method of initialization is preferable.

III. .Demonstration Of Standard Algorithm

Following steps shows the demonstration of *k*-means algorithm:

1.) *k* initial "means" are randomly generated within the data domain.

2.) *k* clusters are created by associating every observation with the nearest mean.

3.) The centroid of each of the *k*-clusters becomes the new mean.

4.) Steps 2 and 3 are repeated until convergence has been reached.

Fig.1. to fig 4 demonstrates above steps:

Input for the algorithm includes data from log file shown in table 1. We consider x-axis as patterns such as p1,p2,p3 etc for pattern 1,pattern 2,pattern 3 etc respectively and then we proceed with the algorithm. The output of this algorithm will consists of clusters of patterns as specified by the user. The input table is shown in table I.

This input is obtained by performing preprocessing steps on the log file shown in fig. 2.

IV. Advantages And Disadvantages

Strength of K-means algorithm:

DOI: 03.LSCS.2013.3.578

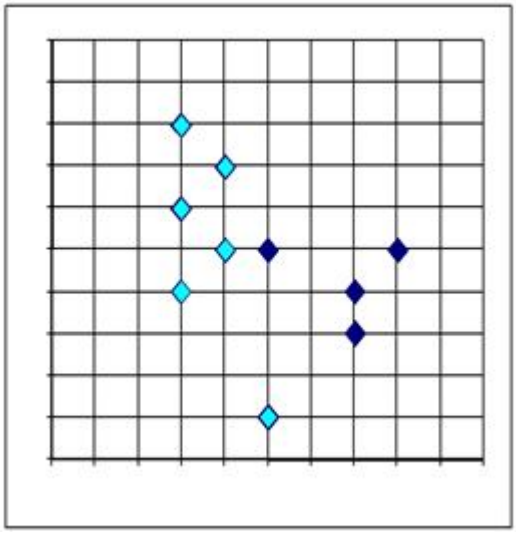

Fig 1. Step1 (k=2)

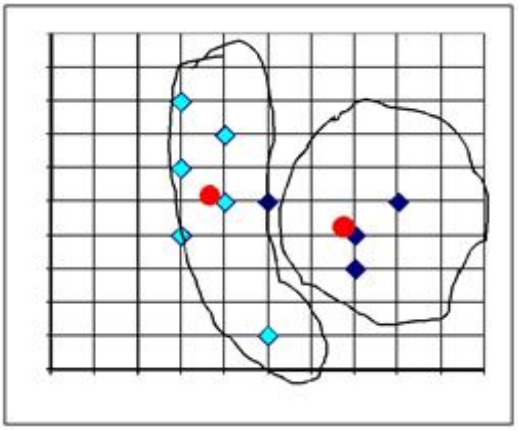

Fig 2.Step 2

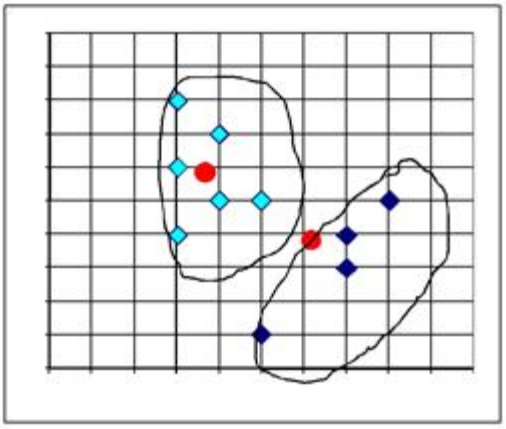

Fig 3.Step 3

The method is relatively scalable.

Efficient in processing large data sets.

Complexity of algorithm is O(nkt),where n is the total number of objects, k is the number of clusters, and t is the number of iterations.

Weakness of K-means algorithm:

Need to specify k, the number of clusters, in advance.

Unable to handle noisy data and outliers.

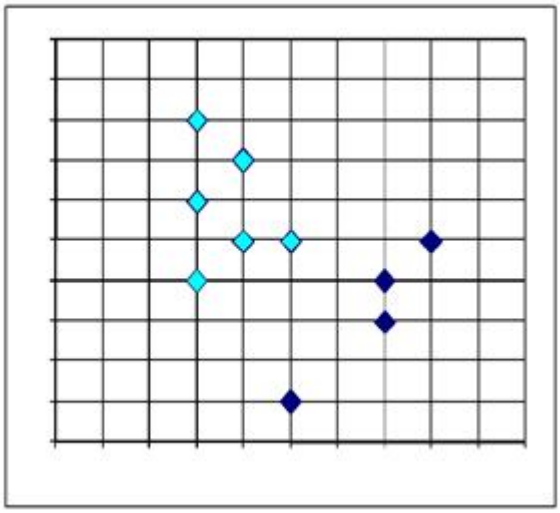

Fig 4. Step 4

TABLE I. INPUT FOR ALFORITHM

Pattern	Time in minutes
P1	120
P2	50
P3	40
P4	12
P5	37

```
login ab@gmail.com 04.02.13 04:08:31
login ab@gmail.com 04.02.13 04:21:30
view ab@gmail.com 04.02.13 04:22:07
upload/pdf/s.pdf
viewout ab@gmail.com 04.02.13 04:22:17
logout ab@gmail.com 04.02.13 04:33:12
login adi4u_2010@yahoo.in 04.03.13
16:14:05
view adi4u 2010@vahoo.in 04.03.13
```

Fig 5. Log data

CONCLUSIONS

An efficient form of k-means algorithm is provided. The algorithm is easy to implement and relatively scalable as complexity of algorithm is O. Only problem with the algorithm is number of clusters have to be specified in advance. The output of algorithm depends upon the no. of clusters specified.

For k=2,we get following clusters:

Cluster1:

P1

DOI: 03.LSCS.2013.3.578

P1
P2
P4

Cluster2:
P2
P2
P3
P5
P1

ACKNOWLEDGEMENT

First, I thank my project Guide and Advisor Mr.Rudresh Shirwaikar, for enlightening me with the subject of data mining throughout. He has always listened to my ideas and supported them and made me think about the subject with a data miner's perspective.

Finally, I am grateful to Shree Rayeshwar Institute of Engineering And Information Technology for supporting me in Information Technology program.

REFERENCES

[1] Hai Dong, Farookh Khadeer Hussain,Elizabeth Chang "A Survey in Traditional Information Retrieval Models", 2008 Second IEEE International Conference on Digital Ecosystems and Technologies.

[2] Determining Usage Patterns on Web Data ,MS Project Report, "Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the Web Data" by Anand Sharma.

[3] Han, J. and Kamber, M. Data Mining: Concepts and Techniques, 2000 (Morgan Kaufmann, San Francisco,California).

[4] J. Breckling, Ed., the Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[5] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[6] E. Schikuta. Grid Clustering: An Efficient Hierarchical Clustering Method for Very Large Data Sets. Proc. 13th Int'l.Conference on Pattern Recognition, 2, 1996802.11, 1997.

[7] K. Alsabti, S. Ranka, and V. Singh. An Efûcient K-MeansClusteringAlgorithm. http://www.cise.uû.edu/ranka/ ,1997.

[8] IEEE Xplore - Research on k-means Clustering Algorithm. http://ieeexplore.ieee.org/xpls/abs_all. jsp?arnumber=5453745

[9] "k-means clustering - Wikipedia, the free encyclopedia."http:/ /en.wikipedia.org/wiki/Kmeans_clustering

[10] Zhang T,Ramakrishnan R,Livny M.BIRCH:An efficient data clustering method for very large databases.In:Jagadish HV,Mumick IS,eds.Proc.of the 1996 ACM SIGMOD Int'l Conf.on Management of Data.Montreal:ACM Press,1996.103-114.

[11] Hinneburg A,Keim D.An efficient approach to clustering in large multimedia databases with noise.In:Agrawal R,Stolorz PE,Piatetsky- Shapiro G,eds.Proc.of the 4th Int'l Conf.on Knowledge Discovery and Data Mining(KDD'98).New York:AAAIPress,1998.58-65

[12] Ding C, He X. K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. In: Proc. of the ACM Symp. on Applied Computing. Nicosia: ACM Press, 2004. 584-589.

[13] Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 1997. 146-151.

[14] Fred ALN, LeitaÞo JMN. Partitional vs hierarchical clustering using a minimum grammar complexity approach.

DOI: 03.LSCS.2013.3.578